HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

BACHELORARBEIT

Herr
Thomas Hache

Matrixdivergenzen bei der Mustererkennung in der Bildanalyse

2010

Fakultät **Mathematik/Naturwissenschaften/Informatik**

BACHELORARBEIT

Matrixdivergenzen bei der Mustererkennung in der Bildanalyse

Autor:
Thomas Hache

Studiengang:
Angewandte Mathematik

Seminargruppe:
Ma07w1-B

Erstprüfer:
Prof. Dr. habil. Th. Villmann

Zweitprüfer:
M.Sc. M. Kästner

Mittweida, 2010

I. Inhaltsverzeichnis

1 Ziele der Arbeit und Motivation

Ziel dieser Arbeit ist es Ähnlichkeiten von Bildern zu analysieren und zu visualisieren. In herkömmlichen Verfahren der Bildanalyse wurden bisher immer die Bildmerkmale in Datenvektoren zusammengefasst und diese dann analysiert. Wir jedoch wollen im Verlauf dieser Arbeit untersuchen, die einzelnen Bilder direkt miteinander zu vergleichen. Man kann deswegen ein Bild auch als Matrix auffassen. Als Ähnlichkeitsmaß wird dann aber nicht der herkömmliche Ansatz benutzt, welcher den euklidischen Abstand einsetzt, sondern wir werden Divergenzen für dieses Problem heranziehen. Dabei gehen wir zunächst auf verschiedene Divergenzklassen ein und stellen diese vor. Da unsere Daten Bilder sind, werden wir speziell für dieses Problem Matrixdivergenzen einsetzen und deren Verhalten analysieren.
Die Arbeit beschränkt sich auf zwei Verfahren der Bildanalyse. Als erstes möchten wir untersuchen, wie sich die Matrixdivergenzen in einem einfachen Visualisierungsprogramm verhalten. Dafür haben wir den t-sne-Algorithmus ausgewählt, der hochdimensionale Daten auf die Ebene projezieren kann. Das Ergebnis wird uns eine Interpretation der Ähnlichkeit zwischen diesen Daten geben. Desweiteren stellen wir noch eine Methode der selbstorganisierenden Karten (SOM) vor. Die Adaption einer SOM benötigt jedoch die Ableitung des Ähnlichkeitsmaß der Objekte, hier Matrizen. Aus diesem Grund werden wir diskutieren, Matrixdivergenzen nach einer Matrix abzuleiten.

2 Vektordivergenzen

Die zu diesem Abschnitt zu Grunde liegende Quelle ist ein Artikel [1], was heißt, dass sich alle Erkenntnisse auf [1] beziehen, wenn nicht anders vermerkt.
Als Einführung zu Divergenzen dient das Kapitel 2, wo der Divergenzbegriff und einige Klassen von Divergenzen vorgestellt wird.

2.1 Einführung

Als Ausgangspunkt betrachten wir eine Wahrscheinlichkeitsdichte $p(x)$. In der Informationstheorie bezeichnet man allgemein

$$-\ln(p(x))$$

als die Information. Wenn wir nun

$$H = \int p(x)\ln(p(x))dx \tag{2.1}$$

bilden, erhalten wir den Erwartungswert der Information, was man auch als Entropie bezeichnet. 2.1 nennt man auch die Shannon-Entropie.

2.2 Divergenzen

Wir betrachten nun zwei verschiedene Wahrscheinlichkeitsdichten $p(x)$ und $q(x)$. Als Divergenz $D(p \parallel q)$ bezeichnet man das Vergleichsmaß für die in den zwei Verteilungen steckende Information. Nach Cichocki und Amari (2010) kann jeder Divergenz eine Entropie zugeordnet werden, wie zum Beispiel der Kullback-Leibler-Divergenz die Shannon-Entropie 2.1:

$$\int p(x)\ln(\frac{p(x)}{q(x)})dx = \int p(x)\ln(p(x)) - p(x)\ln(q(x))$$

Dabei können die Verteilungen p und q für diskrete, als auch für kontinuierliche Werte in die Divergenz eingehen.

$$D_{KL}(p \parallel q) = \sum p_i \ln(\frac{p_i}{q_i}) = \int p(x) \ln(\frac{p(x)}{q(x)}) dx \tag{2.2}$$

Allgemein sind folgende Anforderungen an eine Divergenz zu stellen:

- $D(p \parallel q) \geq 0$
- $D(p \parallel q) = 0 \Leftrightarrow p = q$
- $D(p \parallel q)$ ist konvex im ersten Argument

Eine Divergenz muss allerdings nicht notwendigerweise symmetrisch sein oder die Dreiecksungleichung erfüllen. Daraus schließen wir: Eine Divergenz ist keine Metrik.

2.3 Klassen von Vektordivergenzen

Im folgenden Abschnitt werden wir eine kleine Auswahl von Divergenzenklassen vorstellen. Der Überblick ist deswegen nicht vollständig.

2.3.1 Bregman-Divergenzen

Wir werden im Kapitel 3 Bregman-Matrixdivergenzen aufstellen. Als Grundlage dazu dient uns jetzt die Betrachtung von einfachen Bregman-Divergenzen für Vektoren im diskreten bzw. Dichten im stetigen Fall.

Definition: Bregman-Divergenz

Gegeben sei eine erzeugende Funktion Φ. Sei Φ zweimal stetig differenzierbar im Sinne von Fréchet, L der Raum der Lebesque-integrierbaren Funktionen. Dann ist
$D_\Phi^B : L \times L \to \mathbb{R}^+, d.h. p, q \in L$ und

$$D_\Phi^B(p \parallel q) = \Phi(p) - \Phi(q) - \frac{\partial \Phi(q)}{\partial q}(p - q) \tag{2.3}$$

wobei $\Phi : L \to L$.
Somit ergibt sich die verallgemeinerte Kullback-Leibler-Divergenz als

$$D_{\Phi_{KL}}^B = \int p(x) \log(\frac{p(x)}{q(x)}) - (p(x) - q(x)) dx$$

für positive Maße und

$$D^B_{\Phi_{KL}} = \int p(x) \log(\frac{p(x)}{q(x)}) dx \tag{2.4}$$

für Dichten.

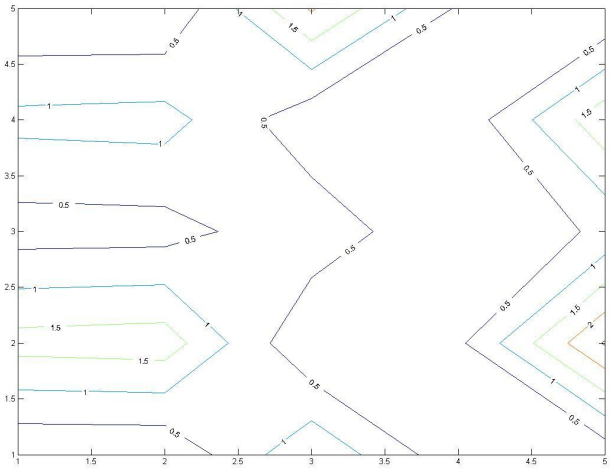

Abbildung 2.1: Höhenlinien der Kullback-Leibler-Divergenz, wobei die Dichten p und $q \in [0,1]$

Als spezielle Bregman-Divergenz ist die Burgdivergenz bzw. Itakura-Saito-Divergenz zu nennen. Diese wird oft in der Bildverarbeitung, Physik und der Soundbearbeitung eingesetzt. Hierbei ist die erzeugende Funktion Φ auf die Burg-Entropie zurückzuführen. Dazu wählen wir

$$\Phi(f) = H^B(f),$$

wobei

$$H^B(f) = -\int \log(f(x)) dx$$

die Burg-Entropie ist und daher sich die Burg-Divergenz als

$$D_{IS}(p \parallel q) = \int [\frac{p}{q} - \log(\frac{p}{q}) - 1] dx \tag{2.5}$$

ergibt.

Desweiteren gibt es in dieser Klasse noch die verallgemeinerte η-Divergenz, wobei $\Phi(f) = f^\eta$ ist. Dadurch ergibt sich:

$$D_\eta(p \parallel q) = \int p^\eta + (\eta - 1)q^\eta - \eta p q^{\eta-1} dx \tag{2.6}$$

Man kann zeigen, dass sich für $\eta = 2$ die quadratische euklidische Distanz ergibt.

Im Folgenden wollen wir auf die Eigenschaften der Bregman-Divergenzen eingehen.

- Die Bregman-Divergenzen sind ausschließlich linear:
  $D^B_{\Phi_1+\lambda\Phi_2}(p \parallel q) = D^B_{\Phi_1} + \lambda D^B_{\Phi_2}$
- D ist invariant gegenüber speziellen affinen Transformationen $\Gamma(q)$
- Es gilt der verallgemeinerte Satz des Pythagoras über p, q, τ:
  $D^B_\Phi(p \parallel \tau) = D^B_\Phi(p \parallel q) + D^B_\Phi(p \parallel \tau) + \frac{\partial\Phi(q)}{\partial q}[p-q] - \frac{\partial\Phi(\tau)}{\partial\tau}[p-q]$
- Die Sensitivität der Bregman-Divergenz ist wie folgt definiert:
  $S(p,q) = \frac{\partial^2 D^B_\Phi(p\|p+\varepsilon q)}{\partial\varepsilon^2}|_{\varepsilon=0} = -q\frac{\partial^2\Phi(p)}{\partial p^2}$

2.3.2 f-Divergenzen

Definition: f-Divergenz

Sei $f \in F = \{g | g : [0,\infty] \to \mathbb{R}, g - konvex, g(1) = 0\}$. Dann ergibt sich die allgemeine f-Divergenz als:

$$D_f(p \parallel q) = \int q \cdot f(\frac{p}{q}) dx \tag{2.7}$$

mit $0 \cdot f(\frac{0}{0}) = 0$ und $0 \cdot f(\frac{a}{0}) = \lim_{x\to 0} x \cdot f(\frac{a}{0}) = \lim_{u\to\infty} a \cdot \frac{f(u)}{u}, a \in \mathbb{R}$

f ist eine erzeugende Funktion und $H_f(p) = -\int f(p)dx$ die f-Entropie.

Eigenschaften:

- $D_f(p \parallel q)$ ist konvex in p und q
- D_f ist invariant bezüglich eines linearen shifts in f:
  $D_f(p \parallel q) = D_{\tilde{f}}(p \parallel q)$ mit $f(x) = \tilde{f}(x) + c(x-1), c \in \mathbb{R}$

- verallgemeinerte Symmetrie:
  duale Funktion $f^*(x) = x \cdot f(\frac{1}{x}) \Rightarrow D_f(p \parallel q) = D_{f^*}(p \parallel q)$
  Folgerung: Sei $f = g + g^* \Rightarrow D_f(p \parallel q) = D_f(q \parallel p)$
- Sei $u = \frac{p}{q}$, p und q Dichten. Dann gilt (Liese/Vajda 1987)
  $0 \leq D_f(p \parallel q) \leq \lim_{u \to 0}(f(u) + f^*(u))$
- Man kann die Annahme über die Konvexität von f aufweichen $\rightarrow$ verallgemeinerte f-Divergenz
  $D_f^G(p \parallel q) = c_f \int p - q + g \cdot f(\frac{p}{q}) dx$, mit $c_f = f'(1) \neq 0$

Als spezieller Vertreter der f-Divergenzen gibt es die Hellinger-Divergenz:

$$D_H(p \parallel q) = \int (\sqrt{p} - \sqrt{q})^2 dx \tag{2.8}$$

mit $f = (\sqrt{u} - 1)^2$ und $u = \frac{p}{q}$

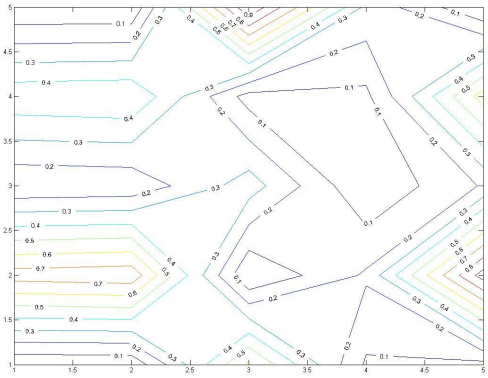

Abbildung 2.2: Höhenlinien der Hellinger-Divergenz, wobei die Dichten p und $q \in [0, 1]$

2.3.3 α-Divergenzen

In Kapitel 3 wollen wir eine α-Divergenz für Matrizen aufstellen. Die Grundlage dafür liefert uns die α-Divergenz für Vektoren.

Die allgemeine α-Divergenz lässt sich wie folgt angeben:

$$D_\alpha(p \parallel q) = \frac{1}{\alpha(\alpha-1)} \int p^\alpha q^{1-\alpha} - \alpha p + (\alpha-1) q dx \tag{2.9}$$

mit $f(u) = u\frac{u^{\alpha-1}-1}{\alpha^2-\alpha} + \frac{1-u}{\alpha}$.

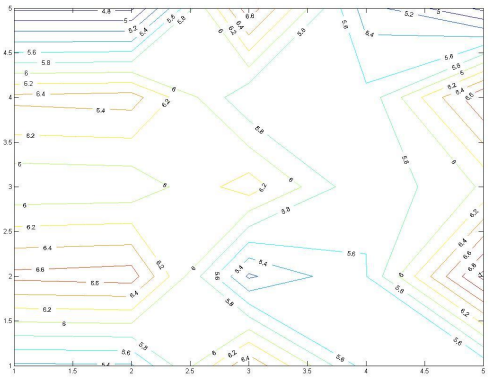

Abbildung 2.3: Höhenlinien der α-Divergenz mit $\alpha = 0.5$, wobei die Dichten p und $q \in [0,1]$

Als spezieller Vertreter der α-Divergenzen ist zum einem die Tsallis-Divergenz zu nennen:

$$D_\alpha^T(p \parallel q) = -\int p \log_\alpha(\frac{p}{q}) dx \tag{2.10}$$

wobei $\log_\alpha(z) = \frac{z^{1-\alpha}-1}{1-\alpha}$.

Zudem kann auch noch die Renyi-Divergenz dazugezählt werden:

$$D_\alpha^R(p \parallel q) = \frac{1}{\alpha-1} \log(\int p^\alpha q^{1-\alpha} dx) \tag{2.11}$$

Eine Besonderheit der α -Divergenz ist, dass gilt:

1. $\lim_{\alpha\to 1} D_\alpha(p \parallel q) = D_{GKL}(p \parallel q)$ und
2. $\lim_{\alpha\to 0} D_\alpha(p \parallel q) = D_{GKL}(q \parallel p)$

Diese Aussage wollen wir nun beweisen.

Beweis:

$$\begin{aligned}\lim_{\alpha\to 1} D_\alpha(p \parallel q) &= \lim_{\alpha\to 1} \frac{\int p^\alpha q^{1-\alpha} - \alpha p + (\alpha-1)q dx}{\alpha(\alpha-1)} \\ &= \lim_{\alpha\to 1} \frac{\int p^\alpha q^1 q^{-\alpha} - \alpha p + (\alpha-1)q dx}{\alpha(\alpha-1)}\end{aligned}$$

Für $\alpha = 1$ erhalten wir den unbestimmten Ausdruck $\frac{0}{0}$. Somit können wir die Regel von L'Hospital anwenden:

$$\begin{aligned}\lim_{\alpha\to 1} D_\alpha(p \parallel q) &= \frac{\frac{\partial}{\partial\alpha}\int q(\frac{p}{q})^\alpha - \alpha p + (\alpha-1)q dx}{\frac{\partial}{\partial\alpha}\alpha(\alpha-1)} \\ &= \lim_{\alpha\to 1} \frac{\int q(\frac{p}{q})^\alpha \ln(\frac{p}{q}) - p + q dx}{2\alpha - 1} \\ &= \int p\log(\frac{p}{q}) - p + q dx \\ &= D_{GKL}(p \parallel q) \quad q.e.d.\end{aligned}$$

Für $\alpha \to 0$ ergibt sich der Beweis analog.

2.3.4 γ-Divergenzen

Eine sehr robuste Klasse, was Ausreißer anbetrifft, wurde von Fujisawa und Eguchi [1] hervorgebracht. Sie heißen γ-Divergenzen, die definiert sind durch

$$\log\left(\frac{(\sum_{i=1}^{n} p_i^{\gamma+1})^{\frac{1}{\gamma(\gamma+1)}} \cdot (\sum_{i=1}^{n} q_i^{\gamma+1})^{\frac{1}{\gamma+1}}}{(\sum_{i=1}^{n} p_i q_i^{\gamma})^{\frac{1}{\gamma}}}\right).$$

Für $\gamma \to 1$ erhalten wir eine, für den weiteren Verlauf wichtige Divergenz. Die Cauchy-Schwarz-Divergenz.

$$D_{CS}(p \parallel q) = \log\left(\frac{\parallel p \parallel \cdot \parallel q \parallel}{< p, q >}\right) \tag{2.12}$$

3 Einschub Funktionale

Da wir Divergenzen über Wahrscheinlichkeitsdichten definiert haben, kann man Vektordivergenzen auch als Funktionale betrachten. Die Idee ist, dass sehr hochdimensionale Datenvektoren als Approximation für stetige Funktionen/positive Maße/Dichten gesehen werden können. Man geht daher von der diskreten Darstellung in die stetige über. Dies erleichtert uns die Gradientenberechnung erheblich, führt aber auf Funktionalableitungen, weil Divergenzen dann dach Funktionen abgeleitet werden müssen. Im Folgenden werden wir einen kleinen Einschub von Definitionen und Funktionalableitungen geben.

3.1 Definitionen und nötige Begriffe

Definition: Funktional [2]

Eine stetige lineare Abbildung zwischen normierten Räumen nennt man einen stetigen Operator. Wenn der Skalare Körper der Bildraum ist, benutzt man anstelle der Bezeichnung Operator die Bezeichnung Funktional.

Definition: Hilbertraum [3]

Ein linearer Raum H mit dem inneren Produkt $< x,y > \in \mathbb{C}$, der bezüglich der durch die Norm $\| x \| = \sqrt{< x,x >}$ induzierten Metrik $d(x,y) = \| x - y \|$ vollständig ist, heißt Hilbertraum.

Definition: Cauchy-Folge [1]

Sei (x_k) eine Folge. Wenn $\forall \varepsilon > 0 \exists n_\varepsilon$, so dass $\forall n,m \geq n_\varepsilon$ gilt $d(x_m,x_n) < \varepsilon$, dann ist (x_k) eine Cauchy-Folge.

Definition: Banachraum [1]

Ein Banachraum ist ein normierter Raum, der vollständig ist. Ein metrischer Raum heißt vollständig, wenn jede Cauchy-Folge konvergiert, der Grenzwert im Raum liegt und die Metrik nicht aus einer Norm stammt.

Bemerkung:

Ein unitärer Raum, der auch Banachraum ist, heißt Hilbertraum [1].

Rieszt'scher Darstellungssatz [4]:

Für einen Hilbertraum H und ein festes $g \in H$ wird durch
$\Lambda : f \mapsto \Lambda f =< f,g >, \forall f \in H$ (*) ein stetiges lineares Funktional auf H mit $\| \Lambda \| = \| f \|$ definiert. Umgekehrt gibt es zu jedem stetigen linearen Funktional $\Lambda \in H'$ genau ein $g \in H$, sodass (*) gilt.

3.2 Funktionalableitungen

1. endlichdimensionale Analogie [1]

Sei V ein Vektorraum und $L : V \to \mathbb{R}$. Desweiteren sei $\mathbf{v} \in V$ und $\mathbf{h} \in V$ ein Richtungsvektor. Dann definieren wir:

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(L(\mathbf{v} + \varepsilon \mathbf{h}) - L(\mathbf{v})) := \frac{dL(\mathbf{v})}{d\mathbf{v}}(\mathbf{h}).$$

Die Existenz des Limes entspricht dabei der Existenz der Ableitung. Die Ableitung ist linear in $[\mathbf{h}]$ und kann durch ein Skalarprodukt ausgedrückt werden.

$$\frac{dL(\mathbf{v})}{d\mathbf{v}} = gradL = \nabla L$$

2. Funktionalableitung [1]

Sei V ein Vektorraum und L ein Funktional, $L : V \to \mathbb{R}$. Ein Punkt in V ist eine Funktion f und eine Richtung in V ist eine Funktion h. Dadurch ergibt sich in Analogie:

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(L(f + \varepsilon h) - L(f)) := \frac{\partial L[f]}{\partial f}[h].$$

$\to$ Frèchet-Ableitung

Sei nun $V = C^\infty(\mathbb{R}^d)$ ein Hilbertraum. Damit definiert sich das Skalarprodukt zweier Funktionen f und g wie folgt:

$$(f,g) = \int f(x)g(x)dx \quad (Lebesque)$$

Rechenregeln für Fréchet-Ableitungen [1]

1. Sei L ein lineares Funktional und $L[f+\varepsilon h]-L[f]=\varepsilon L[h]$. Dann folgt daraus, dass $\frac{\partial L}{\partial f}[h]=L[h]$ und $\frac{\partial L}{\partial f}=L$. Wenn nun $L[f]=\int f(x)g(x)dx=\Lambda[f]$, dann ist $\frac{\partial L}{\partial f}=\int f\cdot g dx$
2. Sei $L[f]=\int F(f(x))dx$ und $F:\mathbb{R}\to\mathbb{R}$, F einmal stetig differenzierbar. Wir betrachten

$$\begin{aligned}\frac{1}{\varepsilon}(L[f+\varepsilon h]-L[f]) &= \frac{1}{\varepsilon}\int F(f(x)+\varepsilon h(hx))-F(f(x))dx\\ &= \int F'(f(x))\cdot h(x)dx.\end{aligned}$$

   Daraus folgt

$$\frac{\partial L[f]}{\partial f}=F'(f).$$

Wir müssen jedoch beachten, dass $\int F'(f(x))\cdot h(x)dx$ bestimmt ist durch den Integralkernel [14] $F'(f(x))=Q(g(x),f(x))$, was wir aber abkürzend als Fréchet-Ableitung $\frac{\partial L[f]}{\partial f}$ schreiben.

Da wir im Folgenden uns mit Matrixdivergenzen beschäfftigen wollen, wird das Thema Funktionalableitungen keine weitere Rolle spielen. Es wäre natürlich möglich, Bilder als Funktionen mit 2 Variablen aufzufassen um auf Funktionale zu schließen. Dieser Aspekt wird aber nicht weiter untersucht und bleibt offen.

4 Matrixdivergenzen

Auf der Grundlage der Vektordivergenzen aus Kapitel 1 wollen wir nun Divergenzen für Matrizen aufstellen um später ein Ähnlichkeitsmaß zwischen zwei Bildern berechnen zu können. Speziell haben wir Bregman-Matrixdivergenzen untersucht, die in Analogie wie 2.3 definiert werden.

4.1 Bregman-Matrixdivergenzen

Im folgenden wird auf Quelle [5] Bezug genommen. Während der Praktikumstätigkeit wurden mehrere Matrixdivergenzen untersucht. Diese benötigt man für die Analyse von Bildern, welche als Datenmatrizen $\mathbf{P}$ gegeben sind. Da wir uns mit Grauwertbildern beschäftigen, nehmen die Elemente p_{ij} nur Werte zwischen 0 und 1 an. Anschaulich betrachtet würde, p_{11} die Graustufe für den Pixel links oben in der Ecke des Bildes darstellen. Je größer die Dimension der Datenmatrix P, desto höher wird die Auflösung des Bildes. Ziel ist es nun mit Hilfe von Matrixdivergenzen ein Ähnlichkeitsmaß für Bilder aufzustellen. Im folgenden wird ein Überblick über die Klasse der Bregman-Matrixdivergenzen gegeben.

Die unten aufgeführten Divergenzen gehören zur Klasse der Bregman-Divergenzen, die von L.M.Bregman 1967 untersucht wurden. Sie sind für alle symmetrischen, positiv definiten $n \times n$- Matrizen geeignet, welche die folgende Eigenwertzerlegung besitzen:

$$\mathbf{P} = \mathbf{V}\Lambda\mathbf{V}^T = \sum_{n}^{i=1} \lambda_i \mathbf{v_i}\mathbf{v_i^T} \tag{4.1}$$

$$\mathbf{Q} = \mathbf{U}\tilde{\Lambda}\mathbf{U^T} = \sum_{n}^{i=1} \tilde{\lambda}_i \mathbf{u_i}\mathbf{u_i^T}$$

Wir haben im Abschnitt Vektordivergenzen die Bregman-Divergenz 2.3 eingeführt. Die allgemeine Bregman-Matrixdivergenz wird nun wie folgt definiert:

$$D_\Phi(\mathbf{P} \parallel \mathbf{Q}) = \Phi(\mathbf{P}) - \Phi(\mathbf{Q}) - tr((\bigtriangledown\Phi(\mathbf{Q}))^T(\mathbf{P}-\mathbf{Q})) \tag{4.2}$$

Mit tr möchten wir die Spur einer Matrix (=Summe aller Hauptdiagonalelemente) abkürzen. Nun müssen wir noch Φ definieren. Wir konstruieren eine konvexe Funktion Φ in Abhängigkeit zum i-ten Eigenwert von $\mathbf{P}$. Wenn man die Eigenwertzerlegung 4.1 betrachtet, sieht man, dass die Spur von $\mathbf{P}$ nur von λ abhängt. Sei $f(\lambda)$ eine monoton

steigende Funktion, dann ist

$$\Phi(\mathbf{P}) = \sum_i f(\lambda_i) = trf(\mathbf{P})$$

eine konvexe Funktion von P. Diese führt zu verschiedenen Bregman-Matrixdivergenzen, wobei im folgenden ein paar Beispiele erläutert werden.

Beipiele:

1. Für $f(\lambda) = \frac{1}{2}\lambda^2$ und $\Phi(\mathbf{P}) = \frac{1}{2}tr(\mathbf{P}^T\mathbf{P})$ ergibt sich dadurch die quadratische Frobeniusnorm oder auch Frobeniusdivergenz:

$$D_F(\mathbf{P} \parallel \mathbf{Q}) = \frac{1}{2} \parallel \mathbf{P} - \mathbf{Q} \parallel_F^2 \tag{4.3}$$

2. Für $f(\lambda) = \lambda ln(\lambda) - \lambda$ und $\Phi(\mathbf{P}) = tr(\mathbf{P}\ln(\mathbf{P}) - \mathbf{P})$ erhalten wir die von-Neumann-Divergenz:

$$D_{vN}(\mathbf{P} \parallel \mathbf{Q}) = tr(\mathbf{P}\ln(\mathbf{P}) - \mathbf{P}\ln(\mathbf{Q}) - \mathbf{P} + \mathbf{Q}) \tag{4.4}$$

3. Für $f(\lambda) = -\ln(\lambda)$ und $\Phi(\mathbf{P}) = -\ln(\det\mathbf{P})$ ergibt sich die Burg-Matrixdivergenz oder Itakura-Saito-Matrixdivergenz:

$$D_{Burg}(\mathbf{P} \parallel \mathbf{Q}) = tr(\mathbf{P}\mathbf{Q}^{-1}) - \ln\det(\mathbf{P}\mathbf{Q}^{-1}) - n \tag{4.5}$$

4. Für $f(\lambda) = \frac{1}{\alpha(\alpha-1)}(\lambda^\alpha - \lambda)$ und $\Phi(\mathbf{P}) = \frac{tr(\mathbf{P}^\alpha - \mathbf{P})}{\alpha^2 - \alpha}$ erhält man die allgemeine Alpha-Divergenz:

$$D_A^{(\alpha)}(\mathbf{P} \parallel \mathbf{Q}) = \frac{1}{\alpha(\alpha-1)} tr(\mathbf{P}^\alpha \mathbf{Q}^{1-\alpha} - \alpha\mathbf{P} + (\alpha-1)\mathbf{Q}) \tag{4.6}$$

Die letzte Divergenz gehört jedoch nicht zu der Klasse der Bregman-Divergenzen. Man kann sie aber sowohl aus der f-Divergenz 2.7, als auch aus der Bregman-Divergenz 2.3 herleiten [5].

Eine weitere wichtige Matrixdivergenz, die ebenfalls keine Bregman-Divergenz ist, ist die Cauchy-Schwarz-Divergenz für Matrizen. Man kann sie mit Hilfe der bekannten Vektordivergenz 2.12 herleiten, indem man beachtet, dass aus $< \mathbf{x}, \mathbf{y} > \Rightarrow tr(\mathbf{A}^T\mathbf{B})$ folgt und insbesondere $\parallel * \parallel = \sqrt{< *, * >}$ ist. Diese Erkenntnisse setzen wir nun in 2.12 ein.

Somit ergibt sich die Cauchy-Schwarz-Divergenz für Matrizen wie folgt:

$$D_{CS}(\mathbf{P} \parallel \mathbf{Q}) = \log(\frac{\sqrt{tr(\mathbf{P}^T\mathbf{P})tr(\mathbf{Q}^T\mathbf{Q})}}{tr(\mathbf{P}^T\mathbf{Q})}) \tag{4.7}$$

Basierend auf den eben aufgeführten Beispielen betrachten wir im kommenden Abschnitt die Ableitungen zu den Matrixdivergenzen.

4.2 Ableitungen von Matrixdivergenzen

Die folgenden Ableitungen wurden mit Hilfe von Quelle [6] erarbeitet. Die Ableitung von Divergenzen ist für die adaptiven Verfahren der Bildanalyse sehr wichtig, denn sie ist die Grundlage für die Berechnung von Gradienten. Diese benötigen wir wiederum, um bei einer SOM den stochastischen Gradientenabstieg in jeder Iteration zu berechnen. Die Matrixdivergenzen aus dem vorherigen Abschnitt wollen wir nun ableiten. Dabei ist für uns nur die Ableitung nach der Matrix $\mathbf{Q} = (q_{ij})$ interessant, da diese das Bild (Prototyp) symbolisiert, welches wir mit dem anderen Bild $\mathbf{P}$ vergleichen wollen.

Zur Ableitung einer Matrixdivergenz nach einer Matrix **Q** benutzen wir folgende Vorgehensweise:

$$(\frac{\partial \mathbf{Q}}{\partial q_{ij}})_{ij} = 1$$

Da wir nach der gesamten Matrix $\mathbf{Q}$ ableiten, betrachten wir alle Ableitungen $\frac{\partial \mathbf{Q}}{\partial q_{ij}}|_{i,j=1,\ldots,n}$ und schreiben diese formal als:

$$\frac{\partial \mathbf{Q}}{\partial \mathbf{Q}} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

4.2.1 Ableitung der Frobeniusdivergenz

Als erstes untersuchen wir die Frobeniusdivergenz. Um die Ableitung herzuleiten, wird wieder die Vorgehensweise $\frac{\partial \mathbf{Q}}{\partial q_{ij}}$ benutzt. Wir untersuchen also

$$\frac{\partial D_F(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}}.$$

Die Frobeniusdivergenz kann man auch als

$$\frac{1}{2}(tr((\mathbf{P}-\mathbf{Q})(\mathbf{P}-\mathbf{Q})^T))$$

schreiben, da die allgemeine Frobeniusnorm

$$\parallel \mathbf{Q} \parallel_F = \sqrt{tr(\mathbf{Q}^T\mathbf{Q})}$$

ist. Wir betrachten also

$$\frac{\partial D_F(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}} = \frac{1}{2}\frac{\partial tr(\mathbf{P}\mathbf{P}^T - \mathbf{P}\mathbf{Q}^T - \mathbf{Q}\mathbf{P}^T - \mathbf{Q}\mathbf{Q}^T)}{\partial q_{ij}}.$$

Die Spur einer Matrix ist linear [13], das heißt $tr(\mathbf{P}-\mathbf{Q}) = tr(\mathbf{P}) - tr(\mathbf{Q})$. Wir beachten dazu, dass $tr(\mathbf{P}\mathbf{Q}^T) = tr(\mathbf{Q}\mathbf{P}^T)$ ist [13]. Dadurch können wir unsere Ableitung in mehrere Einzelterme zerlegen.

$$\frac{\partial D_F(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}} = \frac{1}{2}(\frac{\partial tr(\mathbf{P}\mathbf{P}^T)}{\partial q_{ij}} - 2\frac{\partial tr(\mathbf{P}\mathbf{Q}^T)}{\partial q_{ij}} + \frac{\partial tr(\mathbf{Q}\mathbf{Q}^T)}{\partial q_{ij}}) \tag{4.8}$$

Nun nehmen wir uns jeden einzelnen Term her und untersuchen ihn genauer. Da wir nach dem Element q_{ij} ableiten, bietet sich eine Summendarstellung von den einzelnen Ausdrücken an.

Als erstes betrachten wir den Ausdruck $\frac{\partial tr(\mathbf{Q}\mathbf{Q}^T)}{\partial q_{ij}}$. Zunächst bilden wir das Produkt $\mathbf{Q}\mathbf{Q}^T$ in Summendarstellung und erhalten

$$(\mathbf{Q}\mathbf{Q}^T)_{ij} = \sum_{k=1}^{n} q_{ik}q_{kj}^T = \sum_{k=1}^{n} q_{ik}q_{jk}.$$

Als nächstes berechnen wir die Spur dieses Produktes, was sich folgendermaßen ergibt:

$$tr(\mathbf{QQ}^T) = \sum_{l=1}^{n} (\sum_{k=1}^{n} q_{ik} q_{jk})_{ll}$$

Die Summe können wir vereinfachen, indem wir diese auf die Hauptdiagonalelemente reduzieren, welche für die Spur benötigt werden.

$$\sum_{l=1}^{n} (\sum_{k=1}^{n} q_{ik} q_{jk})_{ll} = \sum_{l=1}^{n} \sum_{k=1}^{n} q_{lk} q_{lk} = \sum_{l=1}^{n} \sum_{k=1}^{n} q_{lk}^2$$

Diese Summe, die wir nun haben, können wir jetz nach den einzelnen Elementen q_{ij} ableiten.

$$(\frac{\partial \sum_{l=1}^{n} \sum_{k=1}^{n} q_{lk}^2}{\partial q_{ij}})_{ij} = 2q_{ij} \tag{4.9}$$

Für den Ausdruck $\frac{\partial tr(\mathbf{PQ}^T)}{\partial q_{ij}}$ benutzen wir die selbe Herangehensweise wie eben. Wir beschreiben zunächst das Produkt in Summendarstellung.

$$(\mathbf{PQ}^T)_{ij} = \sum_{k=1}^{n} p_{ik} q_{kj}^T = \sum_{k=1}^{n} p_{ik} q_{jk}$$

Danach bilden wir wieder die Spur über dieses Matrixprodukt.

$$tr(\mathbf{PQ}^T) = \sum_{l=1}^{n} (\sum_{k=1}^{n} p_{ik} q_{jk})_{ll}$$

Wir reduzieren wieder die Summe auf die Hauptdiagonalelemente und erhalten

$$\sum_{l=1}^{n} (\sum_{k=1}^{n} p_{ik} q_{jk})_{ll} = \sum_{l=1}^{n} \sum_{k=1}^{n} p_{lk} q_{lk}.$$

Diese Summe können wir nun wieder nach den einzelnen Elementen q_{ij} ableiten.

$$(\frac{\partial \sum_{l=1}^{n}\sum_{k=1}^{n} p_{lk}q_{lk}}{\partial q_{ij}})_{ij} = p_{ij} \tag{4.10}$$

Der letzte Term $\frac{\partial tr(\mathbf{P}\mathbf{P}^T)}{\partial q_{ij}}$ verschwindet, da die Funktion, die wir nach dem Element q_{ij} ableiten, nicht von der Matrix **Q** abhängt.

Zum Schluss setzten wir 4.9 und 4.10 in 4.8 ein und erhalten:

$$\begin{aligned}(\frac{\partial D_F(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}})_{ij} &= \frac{1}{2}(-2p_{ij}+2q_{ij}) \\ &= q_{ij} - p_{ij}\end{aligned}$$

Es ergibt sich schließlich die Ableitung

$$\frac{\partial D_F(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}} = \mathbf{Q} - \mathbf{P} \tag{4.11}$$

4.2.2 Ableitung der Cauchy-Schwarz-Divergenz für Matrizen

Eine weitere wichtige Ableitung ist die der Cauchy-Schwarz-Divergenz für Matrizen. Wir betrachten also folgendes Problem:

$$\frac{\partial D_{CS}(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}} = \frac{\partial \log(\frac{\sqrt{tr(\mathbf{P}^T\mathbf{P})tr(\mathbf{Q}^T\mathbf{Q})}}{tr(\mathbf{P}^T\mathbf{Q})})}{\partial q_{ij}}$$

Als erstes, um den Ausdruck zu vereinfachen, wenden wir die Logarithmengesezte an, sodass wir wieder folgende Zerlegung in Einzelterme erhalten:

$$\frac{\partial D_{CS}(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}} = \frac{\partial \frac{1}{2}\log(tr(\mathbf{P}^T\mathbf{P}))}{\partial q_{ij}} + \frac{\partial \frac{1}{2}\log(tr(\mathbf{Q}^T\mathbf{Q}))}{\partial q_{ij}} - \frac{\partial \log(tr(\mathbf{P}^T\mathbf{Q}))}{\partial q_{ij}} \tag{4.12}$$

Interessant sind nur die letzten zwei Ausdrücke, da der erste Term nicht von **Q** abhängt und somit dessen Ableitungen nach den Elementen q_{ij} verschwinden. Die Idee ist nun die Kettenregel anzuwenden. Da die Spur einer Matrix eine reelle Zahl zurück gibt, können wir somit zum Beispiel den Ausdruck $\frac{1}{2}\log(tr(\mathbf{Q}^T\mathbf{Q}))$ als reelle Funktion auffassen,

bei denen die Kettenregel definiert ist. Dadurch ergibt sich folgende Ableitung:

$$\frac{\partial \frac{1}{2}\log(tr(\mathbf{Q}^T\mathbf{Q}))}{\partial q_{ij}} = \frac{1}{2tr(\mathbf{Q}^T\mathbf{Q})}\frac{\partial tr(\mathbf{Q}^T\mathbf{Q})}{\partial q_{ij}}$$

Für den Ausdruck $\frac{\partial tr(\mathbf{Q}^T\mathbf{Q})}{\partial q_{ij}}$ verweisen wir auf 4.9. Somit erhält man:

$$\frac{\partial \log(tr(\mathbf{Q}^T\mathbf{Q}))}{\partial q_{ij}} = \frac{\mathbf{Q}}{tr(\mathbf{Q}^T\mathbf{Q})} \tag{4.13}$$

Zuletzt betrachten wir den Term $\frac{\partial \log(tr(\mathbf{P}^T\mathbf{Q}))}{\partial q_{ij}}$. Auch hier wenden wir wieder die Kettenregel an und erhalten:

$$\frac{\partial \log(tr(\mathbf{P}^T\mathbf{Q}))}{\partial q_{ij}} = \frac{1}{tr(\mathbf{P}^T\mathbf{Q})}\frac{\partial tr(\mathbf{P}^T\mathbf{Q})}{\partial q_{ij}}$$

Den Ausdruck $\frac{\partial tr(\mathbf{P}^T\mathbf{Q})}{\partial q_{ij}}$ behandeln wir mit der Erkenntnis von 4.10. Dadurch ergibt sich folgende Ableitung:

$$\frac{\partial \log(tr(\mathbf{P}^T\mathbf{Q}))}{\partial q_{ij}} = \frac{\mathbf{P}}{tr(\mathbf{P}^T\mathbf{Q})} \tag{4.14}$$

Zum Schluss setzten wir 4.13 und 4.14 in 4.12 ein und erhalten:

$$\frac{\partial D_{CS}(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}} = \frac{\mathbf{Q}}{tr(\mathbf{Q}^T\mathbf{Q})} - \frac{\mathbf{P}}{tr(\mathbf{P}^T\mathbf{Q})} \tag{4.15}$$

4.2.3 Weitere Ableitungen

Bei den restlichen Matrixdivergenzen stoßen wir auf mehrere Probleme. Die vollständige Angabe der Ableitung der restlichen Matrixdivergenzen stellt für uns keinen praktischen Nutzen dar, weswegen wir nur die formale Herangehensweise erläutern werden. Zunächst muss man diskutieren wie Potenzen von Matrizen und der Logarithmus einer Matrix abgeleitet wird, was bei der Alpha-, Burg- sowie der von Neumann-Divergenz auftritt. Wir wissen, um die Bregman-Matrixdivergenzen anwenden zu können, müssen unsere Bilder symmetrisch, positiv definit und quadratisch sein. Diese Forderungen sichern wir, in dem für jedes Bild $\mathbf{X}$ eine Substitution durchgeführt wird der Art, dass $\mathbf{Q} = \mathbf{X}^T\mathbf{X}$.

Von der neu entstandenen Matrix $\mathbf{Q}$ wissen wir, dass sie diagonalisierbar ist. Auf Grund dessen lässt sich zur Bildung des Logarithmus von $\mathbf{Q}$, folgender Algorithmus anwenden [12]:

Algorithmus zur Logarithmusbestimmung:

1. Berechne die Eigenvektoren der Matrix $\mathbf{Q}$ und setze diese in die Matrix $\mathbf{V}$ (jede Spalte von $\mathbf{V}$ ist Eigenvektor von $\mathbf{Q}$.
2. Bilde $\mathbf{V}^{-1}$ unter Beachtung von $\mathbf{V}^{-1} = \mathbf{V}^T$, da **Q** symmetrisch ist, und setze $\mathbf{A} = \mathbf{V}^T\mathbf{Q}\mathbf{V}$.
3. $\mathbf{A}$ ist eine Diagonalmatrix, deren Elemente auf der Hauptdiagonalen Eigenwerte von $\mathbf{Q}$ sind. Bilde nun $\ln(\mathbf{Q}) = \mathbf{V}^T ln(\mathbf{A})\mathbf{V}$.

Nachdem wir nun $\ln(\mathbf{Q})$ berechnen können, betrachten wir den Ausdruck $\frac{\partial ln(\mathbf{Q})}{\partial q_{ij}}$. Man wendet zuerst den Algorithmus von oben an und schreibt dies in einer Summendarstellung. Dies sieht dann wie folgt aus:

$$(\mathbf{A})_{ij} = (\mathbf{V}^T\mathbf{Q}\mathbf{V})_{ij} = \sum_{l=1}^{n}\sum_{k=1}^{n} v_{ik}^T q_{kj} v_{lj}$$

Von der Diagonalmatrix $\mathbf{A}$ kann man nun elementweise den Logarithmus bilden. Danach transformiert man diese Matrix wieder über Orthoganle Matrizen zurück und erhält

$$ln(\mathbf{Q}) = \mathbf{V}^T ln(\mathbf{A})\mathbf{V}.$$

Das erhaltene Resultat stellen wir jetzt in einer Summe dar.

$$(ln(\mathbf{Q}))_{xy} = \sum_{s=1}^{n}\sum_{t=1}^{n} v_{xt}^T ln(\sum_{l=1}^{n}\sum_{k=1}^{n} v_{xk}^T q_{ky} v_{ly}) v_{sy}$$

Diese Summe können wir nun nach den Elementen q_{xy} ableiten. Dabei beachten wir die Kettenregel und dass $(\ln(x))' = \frac{1}{x}$ ist und wir erhalten

$$\frac{\partial ln(\mathbf{Q})}{\partial q_{xy}} = \sum_{s=1}^{n}\sum_{t=1}^{n} v_{xt}^T \frac{\sum_{l=1}^{n} v_{xx}^T v_{ly}}{\sum_{l=1}^{n}\sum_{k=1}^{n} v_{xk}^T q_{ky} v_{ly}} v_{sy}.$$

Die Letzte Frage ist nun, wie man eine Potenz einer Matrix nach einem Matrixelement ableitet, was speziell für die Alpha-Divergenz für Matrizen 4.6 der Fall ist. Das heißt wir betrachten im Allgemeinen folgendes Problem:

$$\frac{\partial \mathbf{Q}^{\alpha}}{\partial q_{ij}},$$

wobei $\alpha \in \mathbb{R}$.

Wir verfolgen eine ähnliche Strategie, wie bei der Berechnung des Logarithmus. Das heißt man diagonalisiert die Matrix $\mathbf{Q}$ zunächst und bildet die Potenz von den Einträgen der Diagonalmatrix [11]. Anschliessend leitet man den gesamten Summenausdruck nach dem Matrixelement q_{ij} ab.

Zunächst führen wir die Diagonalisierung der Matrix $\mathbf{Q}$ durch.

$$(\mathbf{A})_{ij} = (\mathbf{V}^T\mathbf{Q}\mathbf{V})_{ij} = \sum_{l=1}^{n}\sum_{k=1}^{n} v_{ik}^T q_{kj} v_{lj}$$

Die Matrix $\mathbf{A}$ ist eine Diagonalmatrix mit den Eigenwerten von $\mathbf{Q}$ auf der Hauptdiagonalen, von der wir die Potenz bilden können. Anschließend transformieren wir die Matrix **A** über orthogonale Matrizen zurück und erhalten

$$\mathbf{Q}^{\alpha} = \mathbf{V}\mathbf{A}^{\alpha}\mathbf{V}^T.$$

Diesen Ausdruck stellen wir nun als Summe dar.

$$(\mathbf{Q}^{\alpha})_{xy} = \sum_{s=1}^{n}\sum_{t=1}^{n} v_{xt}\Big(\sum_{l=1}^{n}\sum_{k=1}^{n} v_{xk}^T q_{ky} v_{ly}\Big)^{\alpha} v_{sy}^T$$

Nun sind wir in der Lage die Ableitung nach den Elementen q_{xy} vorzunehmen.

$$\frac{\partial \mathbf{Q}^{\alpha}}{\partial q_{xy}} = \sum_{s=1}^{n}\sum_{t=1}^{n} v_{xt}\Big(\alpha\Big(\sum_{l=1}^{n}\sum_{k=1}^{n} v_{xk}^T q_{ky} v_{ly}\Big)^{\alpha-1}\sum_{l=1}^{n} v_{xx}^T v_{ly}\Big) v_{sy}^T$$

Man sieht auch hier wieder, dass es sich im Allgemeinen um eine sehr komplexe Ableitung handelt. Für unsere Alpha-Divergenz für Matrizen gibt es jedoch einen Sonderfall, wenn $\alpha = 2$. In dieser Situation ergibt sich

$$D_A^{(2)}(\mathbf{P} \parallel \mathbf{Q}) = \frac{1}{2} tr(\mathbf{P}^2\mathbf{Q}^{-1} - 2\mathbf{P} + \mathbf{Q})$$

Wir nutzen nun die Linearität [13] der Spur aus und erhalten

$$D_A^{(2)}(\mathbf{P} \parallel \mathbf{Q}) = \frac{1}{2} tr(\mathbf{P}^2\mathbf{Q}^{-1}) - tr(\mathbf{P}) + \frac{1}{2} tr(\mathbf{Q})$$

Nun werden wir die Divergenz ableiten, wodurch sich

$$(\frac{\partial D_A^{(2)}(\mathbf{P} \parallel \mathbf{Q})}{\partial q_{ij}})_{ij} = \frac{\partial \frac{1}{2} tr(\mathbf{P}^2\mathbf{Q}^{-1})}{\partial q_{ij}} - \frac{tr(\mathbf{P})}{\partial q_{ij}} + \frac{\frac{1}{2} tr(\mathbf{Q})}{\partial q_{ij}}$$

ergibt. Der Term $\frac{tr(\mathbf{P})}{\partial q_{ij}}$ verschwindet, da keine Abhängigkeit von $\mathbf{Q}$ vorhanden ist. Nach Quelle [6] ergibt sich die Ableitung der restlichen Terme wie folgt:

$$\frac{\partial D_A^{(2)}(\mathbf{P} \parallel \mathbf{Q})}{\partial \mathbf{Q}} = \frac{1}{2}(-(\mathbf{Q}^{-1})^T(\mathbf{P}^2)^T(\mathbf{Q}^{-1})^T + \mathbf{I}),$$

wobei $\mathbf{I}$ die Einheitsmatrix darstellt. Allerdings erweißt sich diese Ableitung als numerisch instabil, da wir in jedem Adaptionsschritt die Inverse von $\mathbf{Q}$ berechnen müssen.

5 Visualisierung mit dem t-sne - Algorithmus

Aufbauend auf der Theorie wurde das Thema Bildanalyse und Clustern von Bildern auch praktisch getestet. Dazu diente ein Datensatz aus dem Internet mit dem Namen hand written digits. Er beinhaltet ein Experiment mit handgeschriebenen Ziffern von 0 bis 9 und umfasst ca. 1600 Grauwertbilder. Der Datensatz bestand aus einer 1593×256 - Matrix. Dazu standen ebenfalls Klassenlabels zur Verfügung, um die einzelnen Bilder identifizieren zu können. Insgesamt gab es 10 Klassen, in denen die Ziffern von 0 bis 9 unterschieden wurden und jede Ziffer in eine der 10 Klassen eingeteilt wurde. Wir haben den Datensatz in Matlab eingelesen und die Labels den Ziffern entsprechend eingefügt. Ziel war es nun die Lage der einzelnen Ziffern zueinander zu analysieren und diese in der Ebene darzustellen. Um die Visualisierung zu gewährleisten, benutzten wir den t-sne - Algorithmus, der als Matlabimplementation zur Verfügung stand (Laurens van der Maaten).

Damit man eine Vortsellung von dem Datensatz der handgeschriebenen Ziffern bekommt, sind hier nun einige Beispiele.

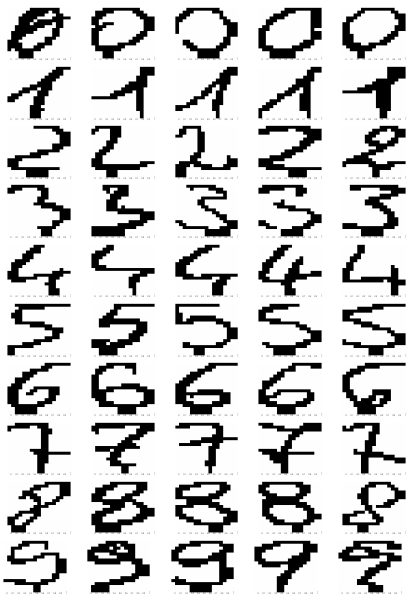

Abbildung 5.1: Ziffern aus dem Datensatz handwritten digits

5.1 Der t-sne - Algorithmus

Der t-sne - Algorithmus [8] liefert eine Abbildung $F : O \to \mathbb{R}^2$, wobei O ein beliebiger Objektraum ist. Üblicherweise betrachtet man $O = \mathbb{R}^n$, wobei $n >> 2$. In unserem Fall ist der Definitionsbereich der Raum der quadratischen, reellen Matrizen, wodurch wir eine Abbildung $F : \mathbb{R}^{n \times n} \to \mathbb{R}^2$ erhalten. Die zu Grunde liegende Kostenfunktion C, welche minimiert werden soll, ist die Kullback-Leibler-Divergenz 2.2:

$$C = \sum_i \sum_j p_{ij} \log(\frac{p_{ij}}{q_{ij}})$$

Hierbei ist p_{ij} die Ähnlichkeit zwischen Bild i und Bild j.

$$p_{ij} = \exp(\frac{-d_{ij}^B}{2\sigma^2})$$

Die d_{ij}^B sind die Ähnlichkeiten zwischen den einzelnen Bildern und wurden mit Hilfe von Divergenzen berechnet. Hierbei wollen wir das unterschiedliche Verhalten von Divergenzen untersuchen. Es ist deswegen offensichtlich, dass für unterschiedliche Divergenzen auch unterschiedliche Visualisierungen entstehen. Die p_{ij}, die Gauss-verteilt sind, wollen wir angleichen mit den q_{ij}, die die Ähnlichkeiten im $\mathbb{R}^2$ repräsentieren und t-verteilt sind.

$$q_{ij} = (1 + d_{ij}^E)^{-1}$$

Hierbei ist d_{ij}^E der quadratische euklidische Abstand.

Durch die Annäherung der p_{ij} an die q_{ij} wird die Kullback-Leibler-Divergenz minimiert (expectation minimization), denn wie man leicht einsieht, wird durch $p_{ij} \approx q_{ij}$ der Logarithmus ≈ 0.

5.2 Analysen der Visualisierungen

Als erstes haben wir eine Visualisierung mit der quadratischen euklidischen Norm durchgeführt. Dabei sind die Ziffern in der Farbe gezeichnet, wie sie hier zu sehen sind: 0, 1, 2, 3, 4, 5, 6, 7, 8(rosa), 9(schwaches grün). Wir erhielten folgendes Ergebnis:

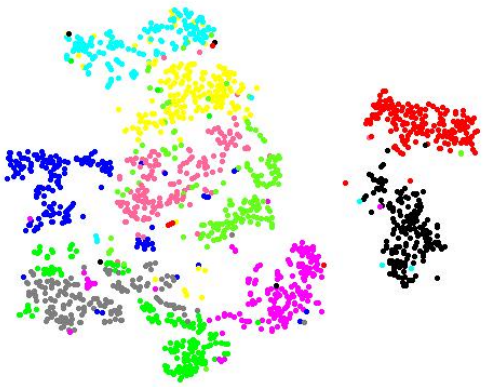

Abbildung 5.2: Visualisierung mit euklidischer Distanz

Man kann hier die deutliche Ähnlichkeit der Ziffer 0 und 6 erkennen, deren Häufungspunkte sehr nahe beieinander liegen. Das selbe kann man bei der 1 und der 7 beobachten, die zum Teil deutlich miteinander verschmolzen sind.

Als nächstes betrachten wir ein Visualisierungsergebnis mit der quadratischen Frobeniusnorm $D_F(\mathbf{P} \parallel \mathbf{Q}) = \frac{1}{2} \parallel \mathbf{P} - \mathbf{Q} \parallel_F^2$:

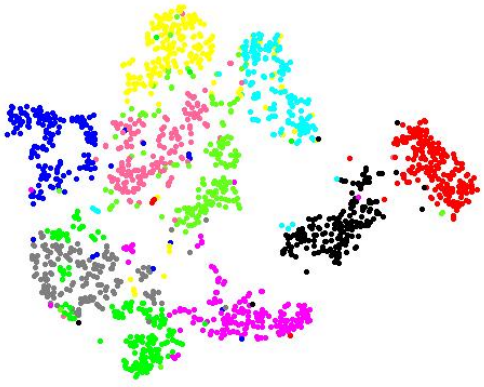

Abbildung 5.3: Visualisierung mit Frobeniusdivergenz

Hier erkennen wir ähnliche Ergebnisse der vorherigen Visualisierung. Die 0 und die 6 verhalten sich näherungsweise genau so wie in der vorhergehenden Rechnung. Das selbe trifft auch für die 1 und die 7 zu.

Zudem haben wir noch eine Visualisierung mit der von-Neumann-Divergenz $D_{vN}(\mathbf{P} \parallel \mathbf{Q}) = tr(\mathbf{P}\ln(\mathbf{P}) - \mathbf{P}\ln(\mathbf{Q}) - \mathbf{P} + \mathbf{Q})$ durchgeführt:

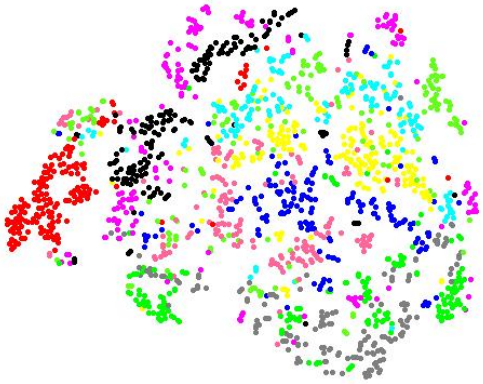

Abbildung 5.4: Visualisierung mit von-Neumann-Divergenz

Dabei erkennen wir einen deutlichen Unterschied zu den anderen zwei Rechnungen. Dies ist darauf zurück zu führen, dass verschiedene Divergenzen auch verschiedene Ergebnisse liefern. Daraus lässt sich schließen, dass nicht jede Divergenz für jedes Problem geeignet ist. Man sollte deswegen untersuchen, welche Divergenz ein problemgerechtes Ergebnis liefert.

Als Letztes stellen wir noch das Ergebnis mit der Cauchy-Schwarz-Divergenz für Matrizen $D_{CS}(\mathbf{P} \parallel \mathbf{Q}) = \log(\frac{\sqrt{tr(\mathbf{P}^T\mathbf{P})tr(\mathbf{Q}^T\mathbf{Q})}}{tr(\mathbf{P}^T\mathbf{Q})})$ vor.

Abbildung 5.5: Visualisierung mit Cauchy-Schwarz-Divergenz für Matrizen

Im Gegensatz zur von Neumann-Divergenz wurden die Ziffern wieder eindeutig getrennt. Man kann wieder die starke Ähnlichkeit der 0 und 6 sowie auch der 1 und 7 erkennen.

5.3 Modifikation der Bilder

Erforscht haben wir auch, wie sich die Matrixdivergenzen verhalten, wenn man die einzelnen Bilder dreht oder spiegelt. Dabei ergab sich folgendes Ergebnis:

Interessanterweise stellte sich die Cauchy-Schwarz-Divergenz bei Drehung und Spiegelung als sehr robust heraus. Bei horizontaler und vertikaler Spieglung wurden die einzelnen Ziffern immer erkannt und separiert in die Ebene geplottet. Zum Schluss transponierten wir noch jedes Bild, was eine Drehung um 90 nach links und eine horizontale Spiegelung bewirkt. Auch hierbei konnte man das Ergebnis problemgerecht einordnen. Im folgenden stellen wir die eben diskutierten Ergebnisse graphisch vor:

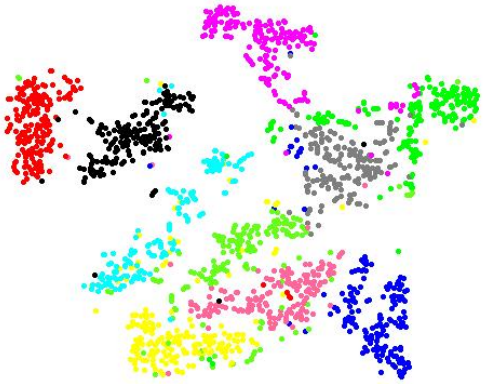

Abbildung 5.6: Visualisierung mit horizontaler Spiegelung

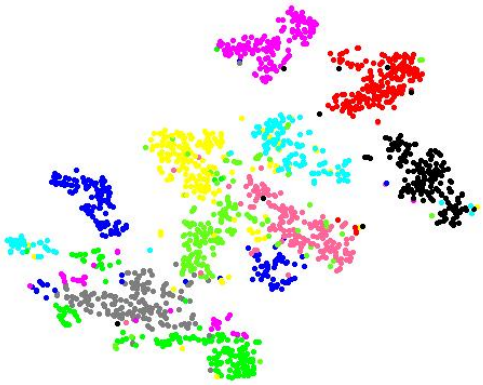

Abbildung 5.7: Visualisierung mit vertikaler Spiegelung

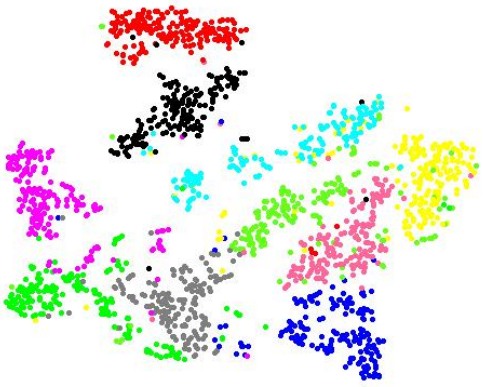

Abbildung 5.8: Visualisierung mit 90 Drehung und horizontaler Spiegelung

Erstaunlicherweise versagte die Frobenius-Divergenz, obwohl sie beim einfachen Visualisieren ein problemgerechtes Ergebnis lieferte. Bei beiden Spiegelungsvarianten und auch bei der Drehung war keine Trennung der einzelnen Ziffern zu erkennen. Die restlichen Divergenzen wurden nicht untersucht, da diese schon für die einfache Visualisierung nicht geeignet waren.

6 Clusteranalyse mit Selbstorganisierenden Karten

In diesem Abschnitt nehmen wir Bezug zu Quelle [9] und [10].

6.1 Einführung zu Selbstorganisierten Karten

Als Einführung betrachten wir zunächst die biologische Motivation. Selbstorganisierende Karten sind eine Untergruppe von künstlichen Neuronalen Netzen. Ihr Funktionsprinzip beruht auf der biologischen Erkenntnis, dass viele Struckturen im Gehirn eine lineare oder planare Topologie aufweisen. Die Signale im Eingangsraum, wie zum Beispiel visuelle oder auditive Reize sind jedoch hochdimensional. Interessant ist nun die Betrachtung, wie diese Reize in planaren Strukturen verarbeitet werden. Durch biologische Untersuchungen kam man zu der Erkenntnis, dass die Eingangssignale bezüglich ihrer Ähnlichkeit abgebildet werden. Das heißt, dass ähnliche Reize nahe beieinander liegen. Es wird also eine Kartierung der Reize vorgenommen. Wenn nun ein Signal vom Körper aufgenommen wird, so werden ausschliesslich die Gebiete der Karte erregt, die dem Signal ähnlich sind. Die mathematische Modellierung dieser Probelmstellung ist auf das Modell von Kohonen zurückzuführen.

6.2 Mathematische Modellierung

Wir werden nun ein zweidimensionales Modell einer Selbstorganisierenden Karte vorstellen, welches auf dem bekannten Modell von Teuvo Kohonen basiert.
Wir betrachten eine Karte A, die sogenannte Neuronenkarte, welche wie folgt definiert ist:

$$A = \{\mathbf{r} = (r_1, r_2) | r_1 \in \{0,1,...,N-1\} \wedge r_2 \in \{0,1,...,N-1\}\}$$

Diese Konstruktion beschreibt ein zweidimensionales Gitter aus $N \times N$ Elementen. Jedes Element $\mathbf{r}$ bekommt einen Gewichtsvektor $\mathbf{q_r}$ aus einem Merkmalraum V zugeordnet, wobei die Komponenten von $\mathbf{q}$ die Eigenschaften der Elemente von A beschreiben. Zwischen den Gewichtsvektoren und den jeweils zugeordneten Elementen von A besteht eine Kopplung, das heisst, es existiert folgende eineindeutige Abbildung

$$\mathbf{e} : A \leftrightarrow V$$

von der Neuronenschicht A in den Merkmalsraum V. Man spricht nun von einer neuronalen Karte.

6.3 Training einer selbstorganisierenden Karte

Als erstes müssen wir beachten, dass unsere Daten Matrizen (Bilder) sind und keine Vektoren. Da die Idee der selbstorganisierenden Karten jedoch auf Vektoren basiert, lösen wir das Problem mit einer einfachen Abbildung $m : \mathbb{R}^n \to \mathbb{R}^{n\times n}$. Diese Abbildung wandelt einen Zeilenvektor $\mathbf{x}$ mit quadratischer Anzahl von Komponenten in eine quadratische Matrix $\mathbf{X}$ um. Man bestimmt zunächst die Dimension n der Matrix, welche die Wurzel aus der Anzahl der Komponenten von $\mathbf{x}$ ist. Dann werden die Komponenten von $\mathbf{x}$ in die Matrix $\mathbf{X}$ geschrieben, wobei aller n Einträge eine neue Zeile bestückt wird. Für unseren Algorithmus benötigen wir natürlich auch eine Abbildung $\bar{m} : \mathbb{R}^{n\times n} \to \mathbb{R}^n$, die eine Matrix $\mathbf{X}$ in einen Vektor $\mathbf{x}$ umwandelt. Dabei gehen wir ähnlich vor wie bei m.
Zu Beginn wählen wir einen Anfangszustand $\mathbf{q}_\mathbf{r}^0$. Die folgende zeitliche Aktualsierung des Gewichtsvektors $\mathbf{q_r}$ wird durch einen iterativen Prozess beschrieben, wobei jede Iteration aus 4 Schritten besteht:

1. Wir wählen ein Bild $\mathbf{P}$ bezüglich einer Wahrscheinlichkeitsverteilung W aus, wobei $W(\mathbf{P})$ in der Neuronenkarten die Reizumgebung beschreibt. $\mathbf{P}$ und $\mathbf{Q_r}$ sind Matrizen von gleicher Dimension.
2. Es wird nun der Ort $\mathbf{s} \in A$ bestimmt (best maching unit) für welchen gilt

$$\mathbf{s} = min_{\mathbf{r}\in A} D(\mathbf{P} \parallel \mathbf{Q_r})$$

3. In diesem Schritt wenden wir eine Funktion $h(\mathbf{r},\mathbf{s},t)$ an, wobei die Wirkung von h auf das Erregungszentrum maximal ist und auf die Nachbarneuronen, in Abhängigkeit der Entfernung zum Erregungszentrum, abnimmt. Wir wählen für h eine einfache Gaussfunktion,

$$h(\mathbf{r},\mathbf{s},t) = \exp(-\frac{(\mathbf{r}-\mathbf{s})^2}{\sigma_h^2(t)}),$$

   wobei $\sigma_h^2(t)$ die Breite der Wirkung der Funktion h, in Abhängigkeit der Zeit t, bestimmt.
4. Zum Schluss wird nun ein Update der Prototypen $\mathbf{Q_r}$ gemäßder Funktion h vorgenommen:

$$\mathbf{Q_r}(t+1) = \mathbf{Q_r}(t) - \alpha(t)h(\mathbf{r},\mathbf{s},t)\frac{\partial D(\mathbf{P} \parallel \mathbf{Q_r})}{\partial \mathbf{Q_r}},$$

wobei $\alpha(t)$ eine streng monoton fallende Funktion in Abhängigkeit der Zeit ist, welche die Lernrate der Karte beschreibt auSSerdem gilt:
$\int_0^\infty \alpha(t)dt = \infty$ und $\int_0^\infty \alpha^2(t)dt < \infty$ für $\alpha(t) > 0$.
Desweiteren legt man $\alpha_{ini} \in [0,1]$ und $\alpha_{fin} \approx \varepsilon > 0$ fest.

Bei dem Training der SOM beschränken wir uns auf ein unüberwachtes Lernen, das heiß t, unsere Prototypen haben während der Lernphase keine Label. Erst nach dem Selbstorganisationsprozess werden den Prototypen Label zugeteilt. Dabei wird der Ausgangsdatensatz zu Hilfe genommen und jeder einzelne Datenpunkt mit den Prototypen nochmals verglichen. Man kann sich dafür den Schritt 1 im obigen Algoritmus vorstellen. Es gibt jedoch zwei verschiedene Modi der Art der Labelvergabe. Man unterscheidet dabei „vote“ und „frequence“. Bei der „vote“-Strategie wird dem Prototyp das Label vergeben, mit dem er am meisten als Gewinnerneuron bestimmt wurde. Man kann aber auch den Modus „frequence“ verwenden. Hierbei bekommt man einen Überblick wann und wie oft ein Prototyp mit irgendeinem Datenpunkt identifiziert wurde.

6.4 Vorüberlegungen

Untersucht haben wir das Lernverhalten der Frobeniusdivergenz und der Cauchy-Schwarz-Divergenz für Matrizen, da aus Erfahrung vom t-sne-Algorithmus diese beiden Matrixdivergenzen vielversprechende Ergebnisse lieferten. Des Weiteren stößt man bei den anderen vorgestellten Matrixdivergenzen bei der Ableitung auf Schwierigkeiten, weswegen diese aus der Betrachtung ausgeschlossen wurden.

Bei den Clusteranalysen haben wir eine Neuronen-Karte vom Format 4×5 genommen, sodass für jede Ziffer im Schnitt 2 Prototypen zur verfügung stehen. Bei der Wahl der Lernrate α muss man jedoch vorsichtig sein. Wenn man sie zu klein wählt, wirkt sich auch das Update im Schritt 4 des Lernprozess sehr gering aus. Andererseits bei hohem α lernt das jeweilige Neuron sehr viel auf einmal. Aus praktischer Erfahrung richtet man sich an einen Wert von $\alpha = 0.1$. Den gesamten Lernprozess haben wir auf 1000 Epochen festgelegt, was bedeutet, dass das Modell mit jedem Datenpunkt 1000-mal lernt. Zur Visualisierung benutzen wir hier das Neuronengitter und die Bilder der Prototypen, die den Mittelwert aller Bilder aus dem Datensatz mit dem selben Label darstellen. Letzteres stellt jedoch für Bregman-Matrixdivergenzen ein Problem dar. Wir erinnern uns an die Forderung, dass die Matrizen symmetrisch, quadratisch und positiv definit sein müssen. Dieses Problem lösten wir gemäß Abschnitt 4.2.3. Damit man die Prototypen besser interpretieren kann, müsste man diese Substitution wieder rückgängig machen.

Die Prototypen **Q** haben die Form

$$\mathbf{Q} = \mathbf{X}^T\mathbf{X}$$

Wobei wir uns nur für $\mathbf{X}$ interessieren, welches die jeweilige Ziffernabbildung repräsentiert. Nun hilft folgende Idee. Man nimmt von jedem Prototyp $\mathbf{Q}$ eine Eigenwertzerlegung vor:

$$\mathbf{Q} = \mathbf{V}^T\mathbf{D}\mathbf{V},$$

wobei **V** die Matrix der Eigenvektoren und **D** eine Diagonalmatrix mit den Eigenwerten von **Q** ist. Die Matrix **D** können wir auch anders schreiben, wodurch sich eine neue Darstellung von $\mathbf{Q}$ ergibt:

$$\mathbf{Q} = \mathbf{V}^T\sqrt{\mathbf{D}}\sqrt{\mathbf{D}}\mathbf{V}.$$

Nun könnte man dieses Produkt folgendermaßen hinschreiben:

$$\mathbf{Q} = (\sqrt{\mathbf{D}}\mathbf{V})^T(\sqrt{\mathbf{D}}\mathbf{V}) \tag{6.1}$$

Diese Aufteilung von **D** ist jedoch nicht eindeutig. Da wir zwischen den beiden Wurzelausdrücken beliebig viele orthogonale Matrizen **O** einschieben können.

$$\mathbf{Q} = \mathbf{V}^T\sqrt{\mathbf{D}}\mathbf{O}^T...\mathbf{O}\sqrt{\mathbf{D}}\mathbf{V}$$

Da aber die Determintante von orthogonalen Matrizen ± 1 ist, bewirkt diese Einschiebung nur eine Drehung bzw. Spiegelung, welche das Bild dennoch interpretierbar macht. Das Problem liegt in der Eigenwertzerlegung, da dort die Eigenvektoren von der Matrix **V** nur bis auf einen Parameter $\alpha \in \mathbb{R}$ bestimmt sind und wir deshalb unendlichviele Eigenvektoren einsetzen können. Wir wissen, dass es eine solche Dartsellung 6.1 gibt, die uns das konkrete Bild des Prototypen liefert, aber wir dies nicht rechnerisch beeinflussen können. Diese Betrachtung bleibt als offenes Problem zurück.

Aufbauend auf der Visualsierung kommen wir nun zu den Ergebnissen.

6.5 Simulation

Wir betrachten zunächst das Ergebnis, bei dem wir mit dem euklidischen Abstand gerechnet haben. Dies dient als Vergleich zu den anderen zwei Divergenzen.

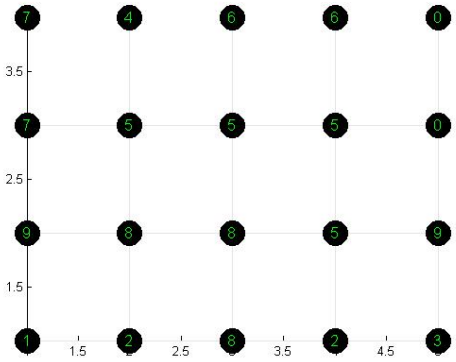

Abbildung 6.1: Lernergebnis mit Euklid. Zu sehen ist ein 4×5-Gitter und das Labeling der Prototypen

Wie erwartet, erkennt man die deutliche Ähnlichkeit der Ziffern 0 und 6. 1 und 7 liegen zwar nicht genau neben einander, aber wenn man das Gitter als Blatt Papier interpretiert und dieses so faltet, dass sich die Ränder berühren, stimmt die Beziehung wieder. Solche Interpretationen muss man durchaus durchführen, da wir nicht genau wissen, wie die Neuronen im Raum verteilt liegen (Problem der Topologieerhaltung).

Als zweites stellen wir das Ergebnis der Frobeniusdivergenz $D_F(\mathbf{P} \parallel \mathbf{Q}) = \frac{1}{2} \parallel \mathbf{P} - \mathbf{Q} \parallel_F^2$ vor:

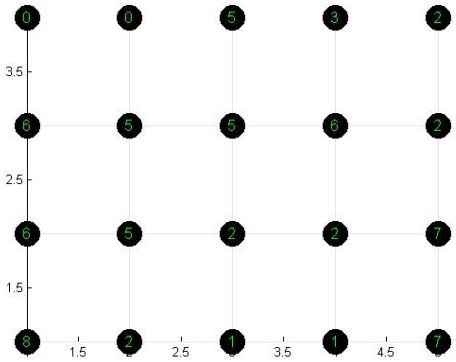

Abbildung 6.2: Lernergebnis mit der Frobeniusdivergenz

Wie auch im euklidischen Ergebnis, sind hier ebenfalls die Ähnlichkeiten von 0 und 6 bzw. von 1 und 7 zu erkennen. Des Weiteren haben wir herausgefunden, dass die Frobeniusdivergenz in diesem speziellen Anwendungsfall auch mit unsymmetrischen Matrizen funktioniert, die nicht positiv definit sind. Diese Erkenntnis erlaubt uns die Prototypen näher zu betrachten. Sehr anschaulich sind hier die 0 und die 6:

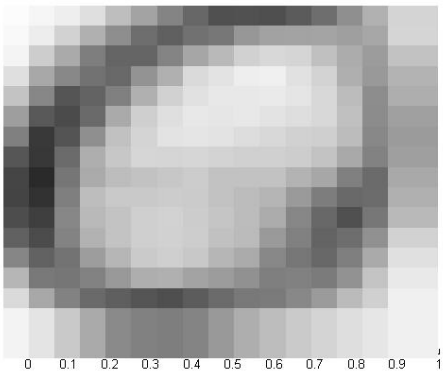

Abbildung 6.3: Prototyp mit dem Label 0

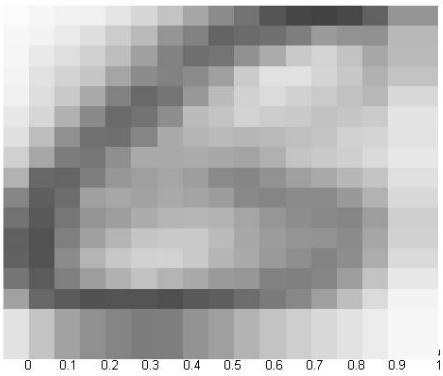

Abbildung 6.4: Prototyp mit dem Label 6

Bei diesen Bildern kann man nochmals erkennen wie ähnlich die beiden Ziffern zueinander sind.

Zum Schluss stellen wir noch das Ergebnis der Cauchy-Schwarz-Divergenz für Matrizen $D_{CS}(\mathbf{P} \parallel \mathbf{Q}) = \log(\frac{\sqrt{tr(\mathbf{P}^T\mathbf{P})tr(\mathbf{Q}^T\mathbf{Q})}}{tr(\mathbf{P}^T\mathbf{Q})})$ vor.

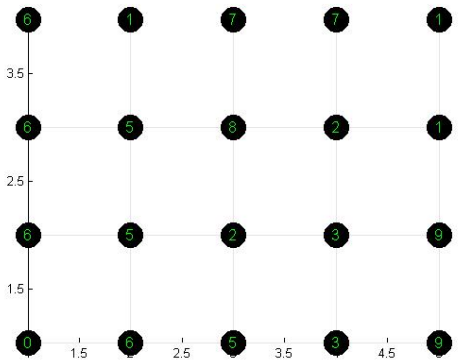

Abbildung 6.5: Lernergebnis mit der Cauchy-Schwarz-Divergenz für Matrizen

Man erkennt auch hier wieder die Beziehung der 0 zur 6 bzw. der 1 zur 7. Da die Cauchy-

Schwarz-Divergenz keine Bregman-Matrixdivergenz ist, gelten die Vorausetzungen der Symmetrie und positiven Definitheit hier nicht. Deshalb können wir auch hier wieder die Prototypen ohne Probleme anschauen. Interessant sind natürlich wie bei der Frobeniusdivergenz die Ziffern 0 und 6.

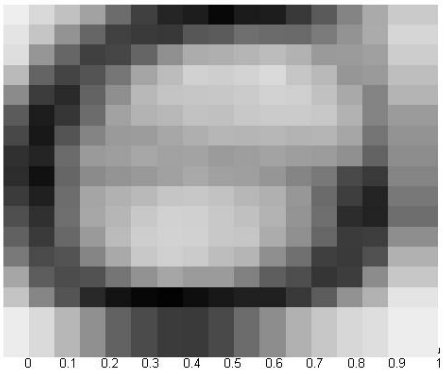

Abbildung 6.6: Prototyp mit dem Label 0

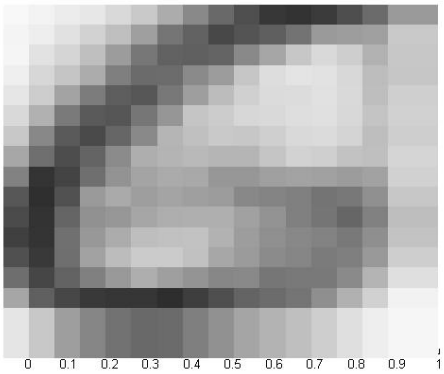

Abbildung 6.7: Prototyp mit dem Label 6

Ähnlich wie bei der Frobeniusdivergenz kann man deutlich die 0 und die 6 identifizieren.

7 Zusammenfassung

Das Ziel der Arbeit wurde erreicht. Wir haben verschiedene Matrixdivergenzen untersucht und herausgefunden, dass jede dieser Divergenzen eine andere Visualisierung hervorruft, was bedeutet, dass jede Divergenz ein anderes Ähnlichkeitsmaß hervorruft. Wir haben gezeigt, dass es möglich ist, Matrixdivergenzen erfolgreich in einem Visualisierungsalgorithmus einzusetzen. Es entstanden problemgerechte Ergebnisse, die zur Präsentation geeignet sind. Desweiteren haben wir Ableitungen von Matrixdivergenzen untersucht. Es stellte sich heraus, dass manche Ableitungen zu schwierig erscheinen, um sie für eine SOM zu benutzen. Je komplizierter die Ableitung wird, desto mehr erhöht sich auch die Rechenzeit für einen Selbstorganisationsprozess. Dafür haben wir gezeigt, dass, bei einer geeigneten Ableitung einer Matrixdivergenz, es auch möglich ist diese in einem Lernprozess mit Neuronen, speziell für den stochastischen Gradientenabstieg, einzusetzen.

8 Literaturverzeichnis

[1] Villmann, Thomas, Haase, Sven: Divergence Based Vector Quantization, 30. September 2010, Mittweida, University of Applied Scinces, Fachbereich Mathematik, Naturwissenschaften, Informatik, Artikel

[2] Werner, Dirk: Funktionalanalysis. 5., erw. Auflage. Springer-Verlag Berlin Heidelberg, Berlin 2005

[3] Bernert, Cordula: Numerik partieller Differentialgleichungen, 2009, Mittweida, University of Applied Scinces, Fachbereich Mathematik, Naturwissenschaften, Informatik, Vorlesung

[4] Darstellungssatz von Riesz, http://mo.mathematik.uni-stuttgart.de/inhalt/aussage/aussage1069/, am 05.06.2010 verfügbar

[5] Cichocki, Andrzej: Nonnegative Matrix and Tensor Factorizations, erste Auflage, John Wiley and Sons, Ltd, 2009, ISBN: 978-0-470-74666-0

[6] Petersen, Kaare Brandt; Pedersen, Michael Syskind: The Matrix Cookbook, Version vom 14.Nov. 2008, http://matrixcookbook.com

[7] Matrixlogarithmus, http://de.wikipedia.org/wiki/Matrixlogarithmus, am 05.06.2010 verfügbar

[8] van der Maaten, Laurens; Hinton, Geoffrey: Visualizing Data using t-SNE, veröffentlicht: Nov. 2008, lvdmaaten@gmail.com

[9] Obermayer, Klaus: Adaptive Neuronale Netze und ihre Anwendung als Modelle der Entwicklung kortikaler Karten, erste Auflage, Technische Universität München, 1992, ISBN: 3-929037-24-6

[10] Martinetz, Thomas: Selbstorganisierende neuronale Netzwerkmodelle zur Bewegungssteuerung, erste Auflage, Technische Universität München, 1992, ISBN: 3-929037-14-9

[11] Griesbach, Ulrich: Lineare Algebra, 2008, Mittweida, University of Applied Scinces, Fachbereich Matehmatik, Natruwissenschaften, Informatik, Vorlesung

[12] Richter, Hans: Zum Logarithmus einer Matrix. In: Archiv der Mathematik. - Haltingen/Baden: 03.02.1950, http://www.springerlink.com/content/l58nj14763542512/, verfügbar am: 05.10.2010

[13] A Matrix-Algebra. URL:<http://www.stat.uni-muenchen.de/ kneib/regressionsbuch/download/matrixanhang.pdf>, verfügbar am 10.10.2010

[14] Kernel (Maschinelles Lernen). URL:<http://de.wikipedia.org/wiki/Kernel>, verfügbar am 18.10.2010

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Mittweida, 13. Dezember 2010